인구통계학적 특성에 따른 협동적필터링 알고리즘의 추천 효율 분석

황성희* · 김영지* · 이미희** · 우용태*

An Analysis of Recommendation Rate for Collaborative Filtering Algorithm based-on Demographic Inforation

Seong-Hee Hwang*, Young-Ji Kim*, Mi-Hee Lee**, Yong-Tae Woo*

요 약

본 논문에서는 고객의 특성을 고려한 최적의 추천시스템을 개발하기 위하여 기존의 인구통계학적 특성에 따른 협동적필터링 기법의 추천 효율을 비교 분석하였다. 비디오에 대한 사용자 평가 값과 예측 값간의 추천 효율에 대한 비교실험을 통하여 상품에 대한 단순한 선호도만을 고려한 기존의 협동적필터링 방법에 의한 추천시스템의 문제점을 개선하여 추천된 상품이나 콘텐츠에 대한 개인별 추천 효율을 향상시키기 위한 모델을 제시하였다. 본 연구 결과를 이용하여 인터넷 비즈니스 분야에서 활발하게 도입되고 있는 eCRM 시스템에서 가장 중요한 요소인 고객들의 인구통계학적인 다양한 특성을 고려한 협동적필터링 기반의 추천시스템을 개발할 수 있으리라 기대한다.

Keywords : 추천알고리즘, 전자상거래, 협동적필터링, eCRM 시스템

I 서 론

전자상거래사이트나 유료로 콘텐츠를 제공하는 사이트가 급격하게 늘어남에 따라 인터넷 비즈니스 분야의 경쟁이 치열하게 전개되고 있다. 특히 인터넷 비즈니스에서 수익 모델 부재로 인해 위기감을 느끼는 전자상거래사이트들은 개인별로 차별화된 일대일 마케팅 전략을 적용하기 위해 eCRM시스템을 활발하게 도입하고 있다. 이러한 상용사이트에서 가장 중요한 이슈는 개인별로 차별화된 정보를 효과적으로 제공하여 고객들의 만족도를 높이기하기 위한 일대일 마케팅 전략으로 이미 아마존과 같은 일부 상용사이트에서 이 기법을 적용하고 있다.

eCRM 시스템 운영을 위한 핵심적인 요소중의 하나는 고객들의 구미에 맞는 상품이나 콘텐츠를 개인별로 권유하기 위한 추천시스템이다. 추천시스템은 데이타마이닝 기법을 이용하여 고객들의 취향과 구매 이력을 분석하여 개인별로 차별화된 정보를 추천하기 위한 자동화된 정보필터링시스템이다. 추천시스템에 관한 기존의 연구는 전통적인 인구통계학적 정보를 이용한 추천(Demographic-based Recommendation) 기법, 정보와 문서간의 키워드 매칭을 통해 정보를 추천하는 내용기반 추천(Content-based Recommendation) 기법, 상품간의 유사도를 계산하여 사용자의 기호에 맞는 상품을 추천하는 항목기반 추천(Item-based Recommendation) 그리고 사용자의 상품에 대한 선호 패턴과 유사한 다른 사용자의 상품에 대한 평가에 기반하여 상품을 추천하는 협동적필터링(Collaborative Filtering)기법 등이 연구되고 있다[Resnick외 4인 1994, Sarwar외 3인 2000, Karypis 2000, Billsus외 1인 2000].

최근에는 상품에 대한 속성을 고려할 필요가 없고, 예상외의 상품에 대한 추천이 가능한 협동적필터링기법에 관한 연구가 활발하게 진행되고 있다. 하지만 이 기법은 상품에 대한 선호도를 중심으로 타겟 사용자와 다른 사용자간의 유사도를 측정하여 유사 집단을 구성하는 관계로 개인별 특성을 고려한 유사 사용자 집단을 구성하기 어렵고, 시스템 구축 초기에는 적용하기 어렵다. 또한 인구통계학적 정보를 이용한 추천기법은 구현 방법이 간단하고 사용자로부터의 피드백 정보가 없이도 추천이 가능하여 시스템 초기 구축 단계나 처음 방문한 고객에 대해서도 적용할 수 있다. 특히 인터넷에서는 다양한 특성을 가진 사용자 계층이 사용하는 관계로 각 계층별로 차별화된 추천 전략이 필요하다. 최근에는 기존의 여러 가지 추천 기법을 결합한 추천시스템에 관한 연구[Claypool외 5인 1999]가 진행되고 있다. 그러나 인구통계학적 특성을 고려한 협동적필터링 기법의 추천 효율을 분석하여 최적의 결합 모델을 제시한 연구는 활발하게 이루어지지않고 있다.

* 창원대학교 전자계산학과, ** 창신대학 정보통신과

본 논문에서는 고객의 집단별 특성을 고려한 최적의 추천시스템을 개발하기 위하여 기존의 인구통계학적 특성에 따른 협동적필터링 기법의 추천 효율을 비교 분석하였다. 제안한 방법은 사용자 세그먼테이션 단계, 세그먼트별 협동적 필터링 분석 단계, 그리고 세그먼트별로 추천 효율을 비교하는 단계로 나뉘어진다.

먼저 사용자 세그먼테이션 단계에서는 고객 집단을 나이나 성별같은 인구통계학적 특성에 따라 세분화하는 과정이다. 세그먼트별 협동적 필터링 분석 단계는 각 세그먼트별로 사용자가 평가한 상품에 대해 고객간의 유사도를 계산하여 유사 사용자 집단을 추출하여 타겟 사용자의 선호도 평가 값을 예측하는 단계이다. 마지막으로 각 세그먼트별로 추천 효율을 비교 분석하여 인구통계학적 특성에 따라 최적의 추천 조건을 분석하였다.

제안된 방법에 대한 효율성을 검정하기 위하여 비디오로 출시된 대표적인 영화에 대한 추천 효율을 비교 분석하였다. 먼저 개인 프로파일 정보에 따라 사용자를 세그먼트하였다. 그리고 타겟 사용자가 속한 세그먼트 내에서 사용자간의 유사도를 계산하여 유사 사용자 집단을 추출한 후 선호도 예측 값이 높은 TOP-N 개의 비디오를 추천하였다. 추천 효율에 대한 성능 비교는 표본 집단에 대하여 평가 값의 일부를 랜덤하게 삭제한 후 협동적필터링 알고리즘에서 예측된 값과의 차이에 대한 평균 값으로 비교하였다. 다양한 사용자 세그먼트별로 협동적필터링 알고리즘을 적용한 결과 협동적필터링 알고리즘만을 적용한 추천 결과와는 상당히 다른 결과를 보였다.

II 추천시스템

2.1 추천시스템의 정의

전자상거래에서 추천시스템이란 사용자 개인의 선호도와 숨어 있는 패턴을 발견하여 개인별로 적절한 상품 정보를 추천해 주는 자동화된 정보 필터링 시스템을 말한다. 즉, 사용자에 대한 인구통계학적 정보, 가장 많이 팔린 상품, 사용자의 구매 패턴 등을 분석하여 개인의 선호도를 데이타마이닝 기법에 의해 분석하여 사용자가 구매하고 싶은 상품을 쉽게 찾을 수 있도록 도와주는 시스템이다.

이러한 추천시스템은 다음과 같은 세 가지 측면에서 전자상거래의 판매를 촉진시키는 효과를 기대할 수 있다. 첫 번째, 추천시스템은 전자상거래 사이트의 방문자에게 고객이 원하는 상품을 쉽게 찾을 수 있도록 도와줌으로써 고객의 구매를 유도할 수 있다. 두 번째, 이미 장바구니에 들어 있는 상품에 기반하여 고객이 구매한 상품 외에 추가적인 상품을 추천하기 위한 교차 판매(Cross-Selling) 전략을 수립할 수 있다. 세 번째, 전자상거래 사이트와 고객간의 가치 있는 관계를 지속적으로 유지하여 고객의 로열티(Loyalty)를 향상시킬 수 있다[Schafer외 1인 1999]. 이미 아마존(www.amazon.com), CDnow(www.Cdnow.com), 예스24(www.yes24.com) 등과 같은 상용사이트에서 추천시스템을 도입하고 있다. 이러한 추천시스템은 전자상거래에 있어서 개인화된 상품이나 콘텐츠를 추천하기 위한 일대일 마케팅 전략을 위한 핵심적인 기법이다. 추천시스템에 관한 기존의 연구 방법은 크게 인구통계학적 정보에 기반한 추천 기법, 내용기반 추천 기법, 항목기반 추천 기법 그리고 협동적필터링 추천 기법 등으로 분류할 수 있다[Resnick외 4인 1994, Sarwar 외 3인 2001, Karypis 2000, 한정기 2001].

2.2 추천시스템에 대한 기존 연구 방법

그 동안 인구통계학적 정보는 특정 유형의 사용자가 선호할 가능성이 있는 상품을 추정하는데 사용되어 왔다. Kruwlich는 사용자 세그먼테이션을 위해 성별, 생활방식, 거주지역 등과 같은 인구통계학적 인자에 대하여 학습 알고리즘을 적용하여 62개의 사용자 세그먼트를 구성하였다[Krulwich 1997]. 또한 Winnow알고리즘은 사용자 세그먼테이션을 위한 학습 방법으로 인자와 가중치 곱의 합으로 사용자간의 유사성을 판정할 수 있다[Pazzani 1999]. 인구통계학적 정보에 의한 추천(Demographic-based Recommendation)기법은 사용자의 성별, 나이, 직업 등과 같은 인구통계학적 요소에 의해 사용자 유형별 특징을 분석하여 상품을 추천하는 방법이다[한정기 2001]. 이 기법은 전통적인 추천시스템의 하나로, 단순한 정보 필터링 기법으로 인해 타겟 마케팅 전략의 하나로 현재까지도 널리 사용되고 있다. 특히 이 기법은 간단하고 사용자로부터의 피드백 정보가 없이도 추천이 가능하여 시스템 초기 구축 단계나 처음 방문한 고객에 대해서도 적용할 수 있지만 추천 효율이 다소 떨어지는 문제가 있다.

내용 기반 추천(Content-based Recommendation) 기법은 개인의 요구나 개인으로부터 입력된 모든 정보와 상품에 포함된 텍스트 정보를 이용하여 필터링하는 방식이다. 이 방법은 논리 연산자(AND, OR, NOT)와 결합된 문자열 등을 검색에 포함시키거나 제거해야할 복잡한 문자열을 가진 텍스트 프로파일을 사용한다. 필터링에 사용되는 텍스트 프로파일은 사용자의 피드백에 의해 자동으로 업데이트된다. 텍스트 프로파일에 대한 업데이팅 기법은 Bayesian Probability, Genetic Algorithm, 또는 Machine Learning 기법 등이 사용된다. 하지만 정보 필터링 과정에서 자신의 프로파일 정보만 이용함으로써 필터링되는 정보가 제한적이다. 또

한 효과적인 추천을 위해서는 상품에 대한 상세한 속성을 포함한 Textual Description 정보가 필요하다[Resnick외 4인 1994]. 이러한 내용기반 추천 방식은 사용자 프로파일을 통해 과거 구매나 추천결과를 쉽게 반영할 수 있는 장점이 있으며 추천 속도도 빠르다. 그러나 상품에 대한 Textual Description 정보의 정확도를 판단하기 어렵고 상품과 사용자가 많은 경우에 효율성이 떨어지는 문제가 있다[Claypool외 5인 1999].

항목 기반 추천(Item-based Recommendation) 기법은 상품간의 유사성을 이용하여 상품을 추천하는 방식이다. Karypis, Badrule 등이 제안한 방식으로 상품간의 관계를 기반으로 하나의 상품에 대한 구매결과로부터 다른 상품에 대한 구매를 유도하기 위한 방법이다[Sarwar외 3인 2001, Karypis 2000]. 이 기법은 코사인 계수와 조건부 확률 등을 이용해 상품간의 유사도를 계산하여 사용자 장바구니에 들어 있는 상품과 유사도가 높은 후보 상품을 추천한다. 이 방식은 상품에 대한 평가 결과가 적은 초기 시스템에서는 사용자간의 유사성을 찾는 협동적필터링 기법보다 유용하지만, 대량의 아이템에 대해서는 비효율적이다. 또한 사용자가 평가한 상품과 유사한 상품 집단을 추천하는 방식을 사용하는 관계로 전자상거래사이트와 같이 상품이 많고 실시간으로 운영되는 환경에서 모든 상품간의 유사도를 실시간으로 계산하기 어렵다. 이 점을 보완하기 위하여 상품간의 유사도 계산을 Pre-computed 모델로 구현할 수 있다[Sarwar외 3인 2001, Karypis 2000].

2.3 협동적필터링 기법

2.3.1 협동적필터링 기법의 개념

최근에 추천시스템에서 가장 널리 사용되는 협동적필터링(Collaborative Filtering) 기법은 타겟 사용자와 유사한 선호도를 가지는 다른 사용자의 상품에 대한 평가를 이용하여 타겟 사용자에게 적절한 상품을 추천하는 방식이다. 먼저 타겟 사용자의 과거 평가 결과를 이용하여 가장 유사한 사용자 집단을 추출하고, 이 유사집단의 사용자가 이미 평가한 상품 중에서 타겟 사용자가 평가하지 않은 상품에 대한 평가 값을 예측하여 추천한다. 이 방식은 사용자 기반(User-based)의 정보 필터링 방식으로, 고객 개인별 추천이 가능하며 예측이 비교적 정확하다. 그리고 상품 자체의 속성에 대한 고려가 필요없고, 예상외의 항목에 대한 추천이 가능하다[Sarwar외 3인 2001, Rafter외 2명 2000].

협동적필터링 기법의 프로세스 개념도는 다음 그림 1과 같다[Sarwar외 3인 2001]. 평가테이블(Rating Table)은 사용자 리스트 $U=\{u_1, u_2, \cdots, u_m\}$와 상품 리스트 $I=\{i_1, i_2, \cdots, i_n\}$로 구성되며, 각 사용자 u_i는 상품들의 평가 값에 대한 리스트 I_{u_i}를 가진다.

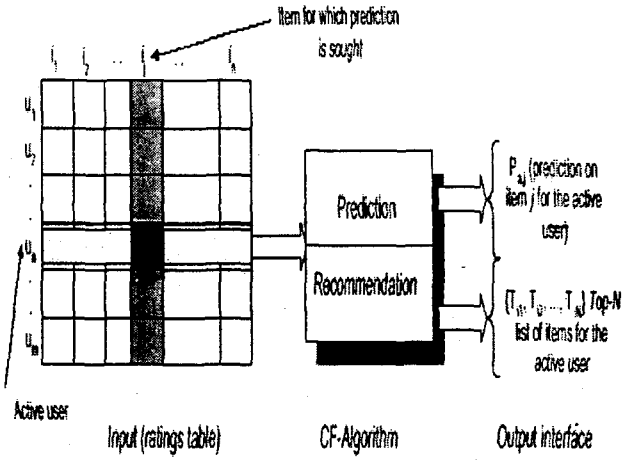

그림 1. 협동적필터링 기법의 프로세스 개념도

평가 테이블은 사용자 리스트 $U=\{u_1, u_2, \cdots, u_m\}$와 상품 리스트 $I=\{i_1, i_2, \cdots, i_n\}$로 구성되며, 각 사용자 u_i는 상품들의 평가 값에 대한 리스트 I_{u_i}를 가진다. 상품에 대한 평가(Rating)는 점수로 표현되거나 구매 기록이나 접속 로그 분석, 웹 마이닝 등으로부터 유추될 수 있다. u_a는 협동적필터링 알고리즘을 수행할 사용자(Active User)이다. 또한 예측(Prediction) $P_{a,j}$는 상품 i_j에 대한 사용자 u_a의 선호도 예측 값이며, $i_j \notin I_{u_a}$이다. 추천리스트 $\{T_{i1}, T_{i2}, \cdots, T_{iN}\}$은 사용자 u_a가 구매하지않은 상품중에서 사용자에게 추천할 상품리스트 $I_r \subset I$이다. 즉, $I_r \bigcap I_{u_a} = \Phi$ 이다.

협동적필터링 기법에 대한 초기 연구는 Tapestry, GroupLens 등이 있으며, Ringo와 Video recommender 등은 E-mail과 웹기반에서 협동적필터링 기법에 의한 추천시스템이다[Karypis 2000]. 그리고 협동적필터링 기법에서 사용되는 유사도 계산은 상관계수를 사용하는 Correlation-base CF(Collaborative Filtering)기법, 사용자와 관계있는 이웃을 찾아내는 K-Nearest Neighbor를 이용한 Memory-based CF기법, 그리고 Bayesian Network, K-Cluster을 사용하는 Model-based CF기법 등이 있다.

2.3.2 Correlation기반의 협동적필터링 기법

협동적필터링 기법중에서 사용자간의 유사도를 측정하기 위해 가장 널리 사용하는 방법은 상관계수를 이용한 Correlation-based CF 기법으로 GroupLens 시스

템에서 처음으로 제안되었다. GroupLens는 Usenet News와 영화에 대한 추천을 위해 협동적필터링을 적용한 자동화된 추천시스템이다. GroupLens에서는 각 기사에 대한 선호도를 데이타베이스에 저장하고 타겟 사용자가 읽지 않은 기사에 대해서 타겟 사용자와 유사하다고 판단되는 다른 사용자의 평가를 기반으로한 상관관계(Correlation)에 의해 평가 결과를 예측한다 [Resnick외 4인 1994, 황병연외 2인 2000].

Correlation-based CF 기법에서는 유사도와 예측 값 계산식을 피어슨 상관계수를 이용해 일반화하였다. 피어슨 상관계수를 이용한 유사도 계산식은 디음 식 (1)과 같다.

$$W_{a,u} = \frac{\sum_{i=1}^{m}(\gamma_{a,j} - \overline{\gamma_a}) * (\gamma_{aj} - \overline{\gamma_u})}{\sigma_a * \sigma_u * n} \quad (1)$$

그리고 피어슨 상관계수를 이용한 예측 값 계산식은 식 (2)와 같다.

$$P_{a,j} = \overline{\gamma_a} + \frac{\sum_{u=1}^{n}(\gamma_{u,i} - \overline{\gamma_u}) * W_{a,u}}{\sum_{u=1}^{n} W_{a,u}} \quad (2)$$

2.3.3 협동적필터링 기법의 문제점

Correlation기반의 협동적필터링 기법이 내포하고 있는 문제점은 크게 초기 평가의 문제, 희소성(Sparsity), Gray Sheep 그리고 상관계수법에 의한 문제점이다 [Claypool외 5인 1999].

초기 평가의 문제점은 시스템 구축 초기에 발생하는 문제로 사용자로부터 충분한 평가와 피드백을 받지 못한 경우 협동적필터링 기법에서는 정확한 추천이 불가능하다는 것이다. 따라서 정확하고 효과적인 추천이 이루어지기 위해서는 충분한 사용자 평가와 피드백 정보가 필요하다.

희소성의 문제는 초기 시스템에서 발생하는 문제와 유사하지만 전자상거래 사이트에서 다양한 상품에 대해서 충분한 평가 결과를 구성하기 어려운 관계로 사용자와 평가로 이루어진 행렬이 희소성을 가진다. 이 문제를 해결하기 위하여 Dimensionality Reduction 기법을 사용하거나, Content-based Software Agent를 사용하여 자동으로 평가를 생성하여 데이타셋의 밀도를 높이는 방법을 사용한다[Sarwar외 4인 2001, Good 외 6인 1999].

Gray Sheep에 대한 문제는 일부 특이한 사용자들에 대한 문제이다. 대부분의 상품에 대해 평균적인 평가를 하여 좋은 것과 싫은 것이 분명하지 않거나, 평가 결과가 일정하지 않은 사용자, 그리고 특이한 성향을 가져 다른 사용자와 유사도를 거의 가지지 않는 사용자는 협동적필터링 기법에서 적절한 추천을 받기 어려운 문제가 있다[Claypool외 5인 1999].

상관계수법에 의한 문제는 사용자의 평가가 적은 경우에는 사용자간 유사도를 구하지 못하여 예측하지 못하게 될 수도 있으며, 예측의 정확도가 떨어진다는 것이다. 또한 사용자의 선호도를 직접적으로 반영하지 못함으로써, 실제 사용자와 유사성이 거의 없고 선호도가 다른 사용자가 이미 평가되어진 아주 적은 수의 상품에 대해서 유사성이 있는 것으로 판단하여 잘못된 추천을 하게 될 가능성이 있다.

III 협동적필터링 기법의 추천 효율 분석

3.1 사용자 정보를 이용한 추천 필요성

인구통계학적 정보를 이용한 전통적인 추천 기법은 사용자로부터 피드백 정보가 없이도 추천이 가능하여 시스템 구축 초기나 처음 고객이 된 사용자에게 유용한 추천 방법이다. 그러나 이 기법은 다른 추천 기법에 비해 추천 효율이 떨어지는 문제점이 있다. 또한 기존의 고객 집단별 마케팅 방식에서 일대일 마케팅 전략으로 전환되고 있는 eCRM 시스템 환경에서 개별화된 마케팅 전략에 적용하기 어렵다.

최근에 상품에 대한 속성을 고려할 필요가 없고, 예상외의 상품 추천이 가능한 협동적필터링 기법에 관한 연구가 활발하게 진행되고 있다. 하지만 협동적필터링 기법은 상품에 대한 선호도에 따라 타겟 사용자와 다른 사용자간의 유사도를 측정하여 상품을 추천하는 관계로 개인별 특성을 고려한 상품 추천이 어렵다.

협동적필터링 기법의 단점을 보완하기 위하여 내용기반 추천기법과 협동적필터링 기법을 결합한 추천시스템[Claypool외 5인 1999]에 대한 연구가 진행되었다. 그러나 기존의 인구통계학적 특성에 따른 협동적필터링 기법의 추천 효율을 체계적으로 분석하여 최적의 결합 방법을 제시한 연구는 아직까지 시도되지 않았다. 따라서 다양한 사용자 계층이 사용하는 인터넷 비즈니스 분야에서 인구통계학적 추천 기법과 협동적필터링 기법을 결합하여 일대일 마케팅이 가능한 추천 모델에 대한 연구가 필요하다.

3.2 사용자 특성에 따른 추천 효율 분석

본 논문에서는 인구통계학적인 특성을 고려한 최적의 추천시스템을 개발하기 위하여 사용자 특성에 따른

협동적필터링 기법의 추천 효율을 비교 분석하였다. 사용자 특성에 따른 협동적필터링 기법의 추천 효율을 분석하는 과정은 크게 사용자 세그먼테이션 단계, 세그먼트별 협동적 필터링 분석 단계 그리고 각 세그먼트별로 추천 효율을 비교 분석하는 단계로 구성된다.

먼저, 고객 집단을 나이나 성별같은 인구통계학적 특성에 따라 세분화하였다. 그리고 타겟 사용자가 속한 세그먼트 내의 사용자들이 평가한 상품에 대해 고객간의 유사도를 계산하여 유사 사용자 집단을 추출하였다. 그리고 유사 사용자 집단으로부터 타겟 사용자에 대한 선호도 평가 값을 예측하여 선호도 예측 값이 높은 TOP-N개의 상품을 추천하였다. 추천 효율에 대한 성능 비교는 표본 집단을 선정하여 평가 점수의 일부를 삭제한 후 협동적필터링 알고리즘에서 예측된 결과 값과의 거리 차에 대한 평균 값으로 비교하였다. 다양한 사용자 세그먼트별로 협동적필터링 알고리즘을 적용한 결과 전체 사용자에 대해 협동적필터링 기법만을 적용한 추천 결과와 상당히 다른 결과를 보였다.

3.3 사용자 정보를 이용한 예측 값 측정

인구통계학적 정보를 이용한 예측 값을 측정하기위해 성별, 나이, 직업 등의 인구통계학적 인자에 가중치를 주어 타겟 사용자와 유사한 사용자를 찾아 예측 값을 정하는 방법을 사용하였다. 사용자간의 유사도를 측정을 위해 Winnow 알고리즘을 사용하였다. Winnow 알고리즘의 적용 과정은 다음과 같다. 먼저, 각 인구통계학적 인자에 대해 가중치를 1로 초기화한다. 다음으로 인자의 집합 $x=\{x_1, x_2, \cdots, x_n\}$ 에서 각각의 인자 x_i에 가중치를 곱한 합을 계산하여 $w_1x_1 + w_2x_2 + \cdots + w_nx_n \geq \tau$ 일 경우 유사도는 1, 그렇지 않은 경우 유사도를 0으로 한다. 만약 잘못된 결과가 나오면 가중치에 2 또는 1/2을 곱하면서 최적의 결과가 나올 때까지 학습한다[Pazzani 1999].

실험에서 사용된 사용자에 대한 프로파일 정보는 인자의 집합 $x=\{$성별,나이,직업,결혼여부$\}$로 표현하였다. 그리고 각 인자에 대한 가중치 w_i를 조정하여 유사 집단을 세그먼트하였다. 그리고 타겟 사용자가 평가하지 않은 비디오에 대한 추천은 유사하다고 판단되는 사용자 집단의 Best-N 리스트를 생성하여 평가 값을 예측할 수 있다.

3.4 협동적필터링기법에 의한 예측값 측정

사용자가 평가하지 않은 상품에 대한 사용자간의 유사도 측정과 예측 값 계산을 위해 Correlation-based CF 기법을 사용하였다. 사용자 U의 비디오 C에 대한 예측 값을 구하기 위해 먼저 식 (1)을 이용해 사용자 U와 다른 사용자와의 유사도를 계산하였다. 그리고 유사성이 있는 사용자들의 평가를 이용해 비디오 C의 예측 값을 식 (2)로 계산하였다. 사용자에 대한 유사도 계산 결과 유사도가 임계치보다 큰 사용자에 대한 평가를 이용해 타겟 사용자에 대한 비디오 i의 예측 값을 계산하였다.

3.5 인구통계학/협동적필터링 결합 모델

기존의 인구통계학적 추천 기법과 협동적 필터링 기법을 결합한 추천 모델에 대한 추천 효율을 분석하였다. 설문조사에서 얻어진 사용자 프로파일의 인구통계학적 정보를 이용하여 사용자를 세그먼테이션 한 후 협동적 필터링 알고리즘을 적용하였다.

먼저 사용자 프로파일의 인구통계학적 인자를 이용해 사용자를 세그먼테이션하고, 협동적필터링 기법을 이용하여 타겟 사용자가 속한 세그먼트 내에서 유사한 사용자를 구한 후 예측 값을 계산하였다. 예를 들어 남자 사용자인 경우 전체 사용자중 남자로 구성된 세그먼트에서 협동적필터링의 상관계수법을 이용하여 유사한 사용자에 대한 예측 값을 구하였다.

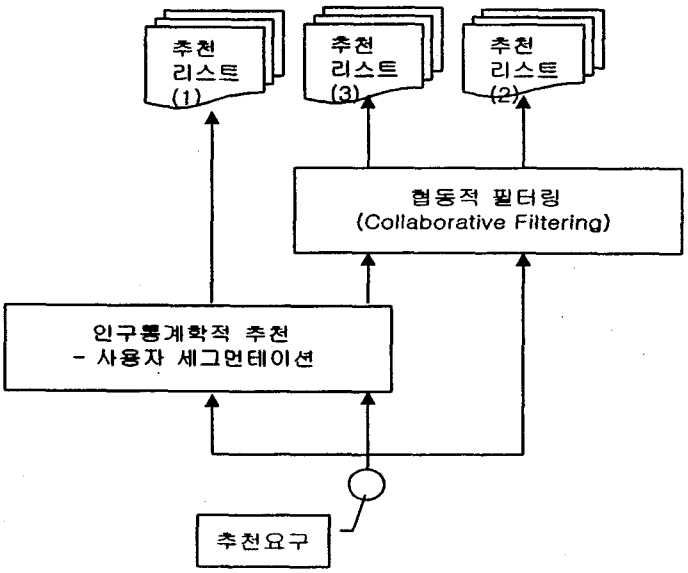

그림 2. 인구통계학적/협동적필터링의 결합 모델

그림 2에서 리스트 (1)은 인구통계학적 추천 기법에 의한 추천 리스트, 추천 리스트 (2)는 협동적필터링에 의한 추천 리스트, 그리고 추천 리스트 (3)은 2가지 기법을 결합한 결과에 의해 생성되는 추천 리스트이다. 실험을 통해 결합 모델의 결과인 추천 리스트 (3)의 추천 효율이 가장 정확함을 알 수 있었다.

IV 실험 결과 및 고찰

4.1 실험 환경

본 논문에서 제안한 인구통계학적 인자를 고려한 추천 기법, 협동적 필터링에 의한 추천 방법 그리고 두 가지를 결합한 모델에 대한 비교 실험을 통하여 각 추천 기법에 대한 추천 효율을 비교 분석하였다. 실험은 ORACLE 8.0.6과 PL/SQL 언어를 사용하여 사용자간의 유사도와 예측 값을 계산하였다.

4.2 실험 결과 및 고찰

제안된 방법에 대한 효율성을 검정하기 위하여 비디오로 출시된 장르별 대표적인 영화에 대한 사용자 평가 값을 분석하였다. 약 235편의 다양한 장르의 비디오에 대하여 고등학생, 대학생, 직장인 등과 같은 서로 다른 계층을 대상으로 자신이 본 비디오에 대한 평가를 오프라인으로 조사하였다. 설문은 성별, 나이, 직업 등과 같은 인구통계학적인 사용자 정보와 각 비디오에 대해 1~5점 사이의 사용자 평가 값을 수집하였다.

Gray Sheep 문제를 해결하기 위하여 평가 값의 표준편차가 일정 임계치이하인 145명에 대한 평가 값을 사전에 제외시켰다. 또한 사용자들의 평가가 거의 이루어지지않은 105편의 비디오에 대한 평가 값도 제외시켰다. 추천 효율에 대한 비교는 사용자 8명을 표본집단으로 선정한 후, 그들의 비디오에 대한 평가 결과의 일부를 무작위로 제거하고 상관계수법에 의한 협동적필터링 알고리즘을 적용하여 예측 값을 계산하였다.

먼저 Winnow 알고리즘에 적용하여 표본집단에 대한 성별과 나이 인자에 동일한 가중치를 부여하여 두 인자가 일치하는 사용자끼리 세그먼테이션하였다. 예측 값은 협동적필터링 기법의 예측 값 계산식 (2)를 통해 계산하였다. 다음 표 1과 그림 3은 추천 기법별로 표본집단에 대한 실제 평가 값과 예측 값과의 차이이며,각 기법간의 추천효율을 의미한다.

표 1. 추천 기법별 추천효율 비교

순번	추천기법	평가/예측값차이
[1]	CF (협동적필터링 기법)	0.94
[2]	CF & 인구통계학(성별)	1.00
[3]	CF & 인구통계학(나이)	0.74
[4]	CF & 인구통계학(성별,나이)	0.59
[5]	인구통계학적 필터링 기법	0.65

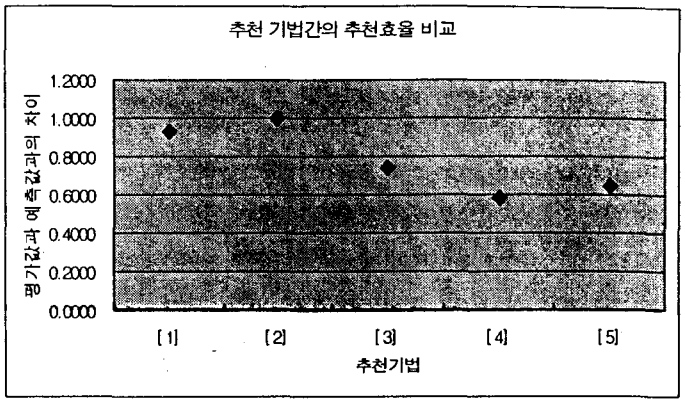

그림 3. 추천 기법별 추천효율 비교

표 1에서 표본집단에 대한 예측 값은 실제 평가 값과 평균적 0.72의 차이를 보였다. 전체 사용자의 예측 값은 실제 평가 값과 평균 0.94의 차이를 보였다.

그리고 인구통계학적 기법과 협동적필터링 기법을 결합한 실험은 인구통계학적 인자에 따라 3가지로 실험하였다. 즉, 성별만 고려한 경우, 나이만 고려한 경우, 그리고 성별과 나이를 모두 고려한 경우에 대해 표본집단에 대한 예측 값은 실제 평가 값과 1.00, 1.00, 0.64의 차이를 보였다. 다음 그림 4는 사용자 세그먼트별 추천 기법을 비교한 결과이다.

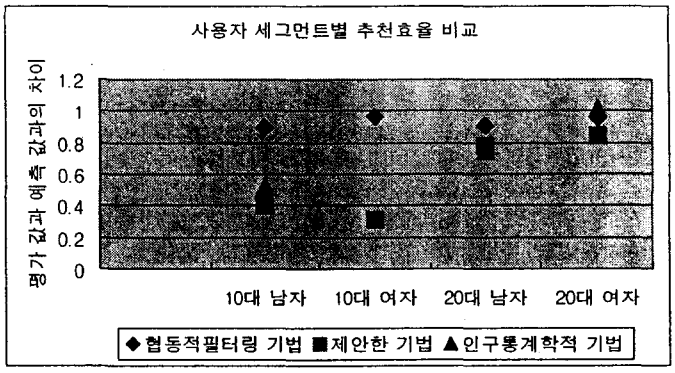

그림 4. 사용자 세그먼트별 추천효율 비교

표 2와 그림 4에서처럼 10대 여자의 경우 인구통계학적 인자에 가장 영향을 많이 받았다. 또 각 세그먼트마다 인구통계학적 인자에 영향을 받는 정도가 차이가 있음을 알 수 있다. 따라서 사용자 개인별 특성을 고려한 효율적인 상품 추천시스템 개발을 위해 인구통계학적 인자가 필수적으로 고려되어야 함을 알 수 있다. 또한 추천 기법간의 적절한 결합에 의해 인구통계학적 추천 기법이나 협동적필터링 기법중 어느 한가지만 사용한 결과보다 추천 효율이 높은 것을 알 수 있다.

V 결 론

본 논문에서는 고객의 집단별 특성을 고려한 최적의 추천시스템을 개발하기 위하여 기존의 인구통계학적 특성에 따른 협동적필터링 기법의 추천 효율을 비교 분석하였다. 제안된 방법에 대한 타당성을 검증하기 위하여 고등학생, 대학생, 직장인 250명에 대해 235편의 비디오에 대한 평가 값을 설문 조사하였다. 응답자의 일부를 표본집단으로 선정한 후, 평가 값과 예측 값간의 차이를 구하여 추천 효율을 비교하였다. 인구통계학적 기법과 협동적필터링 기법을 결합한 모델에 대한 비교 실험은 성별만 고려한 경우, 나이만 고려한 경우, 그리고 성별과 나이를 모두 고려한 경우에 대해 실험하였다. 실험 결과, 표본집단에 대한 예측 값은 실제 평가 값과 1.00, 1.00, 0.64 정도의 차이를 보였다. 특히 10대 여자의 경우 인구통계학적 인자에 가장 영향을 많이 받았고, 각 세그먼트마다 인구통계학적 인자에 영향을 받는 정도가 뚜렷한 차이를 보였다.

결과적으로 사용자 개인별 특성을 고려한 효율적인 상품 추천을 위해 인구통계학적 정보가 필수적으로 고려되어야 함을 알 수 있었다. 앞으로 본 연구 결과를 이용하여 상품에 대한 단순한 선호도만을 고려한 기존의 협동적필터링 기법에 의한 추천시스템의 문제점을 개선하여 추천된 상품이나 콘텐츠에 대한 개인별 추천 효율을 향상시킬 수 있으리라 기대한다. 또한 본 모델은 사용자의 평가 결과가 적은 초기 시스템 환경에서도 추천 효과를 높일 수 있다. 특히 인터넷 비즈니스 분야에서 활발하게 도입되고 있는 eCRM 시스템에서 가장 중요한 요소인 고객들의 인구통계학적인 다양한 특성을 고려한 협동적필터링 기반의 추천시스템을 개발할 수 있으리라 기대된다.

참고문헌

한정기, "개인화(Personalization)의 핵심 기술," http://personalization.co.kr/column[010319].htm, 2001

황병연, 김대겸, 양준호 등, "중·소형 전자상거래를 위한 매칭 에이전트 시스템 개발," 산·학·연 공동기술개발사업 최종 연구보고서, 2000

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," Accepted for publication at the WWW Conference. 2001

Berson, A., Smith S., Thearling, K., "Building Data Mining Applications for CRM," Mc Graw Hill, Chapter 6, 7, 13, 2000

Billsus, D. & Pazzani, M. "Learning Collaborative Information Filters," Proc. of the International Conference on Machine Learning, 1998, pp.46-53

George Karypis, "Evaluation of Item-Based Top-N Recommendation Algorithms," Technical Report CS-TR-00-46, Computer Science Dept., Universiry of Minnesota, 2000.

J. Ben Schafer, Joseph Konstan, John Riedl, "Recommender Systems in E-Commerce," ACM Conference on Electronic Commerce, 1999, pp.158-166

Jonathan L. Herlocker, Josep A.Konstan, "Explaining Collaborative Filtering Recommendations," Proc. of the ACM Conf. on Computer Supported Cooperative Work, 2000, pp.115-152

Krulwich, B., "LIFESTYLE FINDER: Intelligent User Profiling Using Large-Scale Demographic Data," Artificial Intelligence Magazine Vol. 18, No.2, 1997, pp.37-45

Mark Claypool, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes and Matthew Sarti, "Combining Content-Based and Collaborative Filters in an Online Newspaper," ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, 1999

Michael J. Pazzani, "A Framework for Collaborative, Content-Based and Demographic Filtering," Artificial Intelligent Review, 1999, pp.394-408

Nathaniel Good, J. Ben Schafer, Joseph A. Konstan, Al Borchers, Badrul Sarwar, Jon Herlocker, and John Riedl, "Combining Collaborative Filtering with Personal Agents for Better Recommendations," AAAI/IAAI, 1999, pp.439-446

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstorm, John Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proc. of the ACM 1994 Conference on Computer Supported Cooperative Work, pp.175-186, 1994

Rachael Rafter, Keith Bradley, Barry Smyth, "Personalised Retrieval for Online Recruitment Services," 22nd Annual Colloquium on IR Research, 2000

Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. "Analysis of Recommender Algorithms for E-Commerce," Proc. of the 2nd ACM E-Commerce Conference (EC'00). Oct., 2000, pp.158-167

Scott Deerwester, Susan Dumais, Goerge Furnas, Thomas Landauer, and Richard Harshman. "Indexing by latent semantic analysis," Journal of the American Society for Information Science, Vol. 41, No.6, 1990, pp.391-407

Upendra Shardanand, Pattie Maes, "Social Information Filtering: Algorithms for Automating "Word of Mouth," Proc. of the CHI-95 Conference. Denver, Colorado, 1995, pp.210-217

"Seminar in Computational Learning: The Exact Learning Model. Lecture11: The on-line model, learning decision lists, and the winnow algorithm," http://www.cs.bgu.ac.il/~beimel/Courses/Learning/learning99.html